Double Reading Reduces Miss Errors in Low Prevalence Search

Melina A. Kunar[1], Derrick G. Watson[1] & Sian Taylor-Phillips[2]

1) Department of Psychology, The University of Warwick, Coventry, CV4 7AL, UK
2) Warwick Medical School, The University of Warwick, Coventry, CV4 7AL, UK

Email: m.a.kunar@warwick.ac.uk

Tel: +44 (0)2476 522133

Running Title: Paired Reading in LP search

Acknowledgements

The authors would like to thank Peter Carr for help with data collection. This work was supported by grants awarded to Melina Kunar from the British Academy and the Leverhulme Trust (SG122252) and the Experimental Psychology Society, UK. Data from these experiments can be found on the Open Science Framework (https://osf.io/2hq9f/).

Abstract

Low Prevalence studies show that people miss a large proportion of targets if they appear rarely. This finding has implications for real-world tasks, such as mammography, where it is important to detect infrequently appearing cancers. We examined whether having people search in pairs in a 'double reading' procedure reduces miss errors in Low Prevalence search compared to when participants search the displays alone. In Experiment 1 pairs of participants searched for a mass in a laboratory mammogram task. Participants either searched the same display together (in the same room) or searched the displays independently (in separate rooms). Experiment 2 further manipulated the reading order so that paired participants either read the mammograms in the same or different orders. The results showed that, although there was no effect of reading order, double reading led to a substantial reduction in miss errors compared to single reading conditions. Furthermore, the reason for the double reading improvement differed across reading environments: when participants read the displays in a shared environment (i.e. in the same room) the improvement occurred due to an increase in sensitivity, however when participants read the display in different rooms the improvement occurred due to a change in response bias.

Key words: Low Prevalence, Mammogram, Visual Search, Double Reading

Public Significance Statement

The present study suggests that having two people search for a low prevalence target, such as a cancer in a mammogram leads to a reduction in the number of cancers that are missed. Furthermore, this two reader-benefit occurs when both readers view the mammograms together (in the same room) or independently (in different rooms).

Introduction

In everyday life, people perform visual search tasks to find a target. Examples of these range from searching for your car keys to search that informs important medical or safety decisions, such as a radiologist searching for a cancer in a mammogram. This latter search has the extra complication that the target will typically appear infrequently – a factor that is important because targets with a low prevalence rate are often missed (Wolfe, Horowitz & Kenner, 2005).

Wolfe, Horowitz and Kenner (2005) first investigated the effect of prevalence rates on search by asking participants to search for a target item (in their experiments an image of a tool) which could either appear with a High Prevalence (HP, e.g., 50% of the time) or with a Low Prevalence (LP, e.g., 1% of the time). The results showed that participants missed a large proportion of targets when the prevalence was low, compared to when the target prevalence was high. In fact, miss error rates increased from 7%, to 16% to 30% as target prevalence rates decreased from 50% to 10% to 1%, respectively. This 'Low Prevalence' Effect has since been found to be extremely robust and difficult to counter (Wolfe et al., 2007) and the results have been replicated many times with a wide range of stimuli (e.g., Wolfe et al. 2007, Rich et al., 2008, Kunar, Rich & Wolfe, 2010, Russell & Kunar, 2012, Van Wert, Horowitz & Wolfe, 2009, Mitroff & Biggs, 2014, Kunar et al., 2017). Typical LP experiments use a target prevalence rate of 2% (e.g., Wolfe et al., 2007, Fleck & Mitroff, 2007, Rich et al., 2008, Kunar et al., 2010, Russell & Kunar, 2012, Van Wert, Horowitz & Wolfe, 2009) although performance with target prevalence rates of up to 10% have also been examined (Wolfe et al., 2005).

Wolfe et al. (2007) investigated whether the LP Effect occurred simply due to participant fatigue or a decrease in how alert participants are across the experiment. By the very nature of looking for a rare target, participants in an LP experiment have to search through and respond to a large number of displays (e.g., often exceeding 1000 trials). In one experiment, Wolfe et al. (2007) asked participants to self-report their subjective feeling of alertness through an LP search task (containing 1,700 trials). They also performed an objective psychomotor vigilance task (PVT) throughout the experiment to measure alertness. Although the results showed that participants' self-rating of alertness declined over the course of the experiment there was no objective decline in PVT performance. Therefore, Wolfe et al. (2007) concluded that the increase in miss errors due to LP was not simply due to participant fatigue.

Fleck and Mitroff (2007) suggested that the LP Effect occurred due to motor errors. That is, when the target was rare, participants became pre-disposed to respond that the target was absent, as this would typically be the correct response. Therefore, even when the target was present, participants would incorrectly press the wrong button out of 'habit' without it being a true reflection of their perceptual ability to see the target. If this were the case, then giving people a chance to self-correct their motor errors should eliminate the LP Effect. Consistent with this, Fleck and Mitroff (2007) found that the LP Effect was removed following the introduction of a self-correct option. Rich et al. (2008) provided similar evidence but only when the search task was easy (using a single feature task). When the search task was more complex the self-correct option did not eliminate the LP Effect. These findings have been replicated in a number of other studies in which it was found that the self-correct option reduces but does not eliminate the LP Effect and that a substantial LP Effect remains even after people have self-corrected their motor responses (Van Wert, et al., 2009, Kunar, et al., 2010, Kunar et al., 2017, Russell & Kunar, 2012, Rich et al., 2008).

Wolfe and Van Wert (2010) proposed a Multiple Decision Model (MDM) to explain the LP Effect (see Figure 1). The MDM demonstrates how multiple factors influence how people search a display and can account for the typically observed large proportion of miss errors at LP. Wolfe and Van Wert (2010) suggested that in a search task the item with the highest activation, based on top-down and bottom-up interactions, would be selected for further processing. The selected item would then undergo a two alternative forced choice (2AFC) decision process to determine whether it was the target or not. If the outcome of the 2AFC determined that it was the target item then the search process would terminate. If not, the item with the next highest activation would be selected for further processing. This would continue until the target item was found or until participants reached a quitting threshold, where they concluded they had searched the display sufficiently to determine that no target was present. Wolfe and Van Wert (2010) proposed that under LP search two parameters of the MDM differ compared to HP search (see also Peltier & Becker, 2016). The first parameter is that under LP search conditions the quitting threshold is lowered so participants are more likely to terminate their search sooner (concluding there is no target) than if the prevalence of the target (and hence the quitting threshold) had been higher. Converging evidence for this change in parameter comes from both behavioural and eye movement studies. For example, across LP studies, reaction times (RTs) are typically shorter than they are at HP (especially for target absent trials e.g., Wolfe et al., 2007, Kunar et al., 2017, Rich et al., 2008, Russell & Kunar, 2012). Furthermore, eye movement studies have shown that in LP search, people often respond 'target absent' before their eyes have fixated the target (Rich et al., 2008, Peltier & Becker, 2016). The second parameter change under LP conditions is a shift in response criteria. That is, at LP people become more conservative in their responses, requiring more evidence before they conclude that a target item is present.

Several studies have used Signal Detection Theory (SDT) to examine potential changes to both response criteria and sensitivity on LP trials (Wolfe et al., 2007, Van Wert, Horowitz & Wolfe, 2009, Wolfe & Van Wert, 2010, Russell & Kunar, 2012). Specifically, SDT (Green and Swets, 1967, Macmillan & Creelman, 2005) allows us to determine whether the high proportion of miss errors in LP search are due to a change in sensitivity (as measured by *d'*), a change in response selection (as measured by *c*) – or a combination of both. A change in *d'*, means that participants became less sensitive in search so that they were less able to distinguish the presence of a target within an image. A change in *c* on the other hand means that the decision criterion had shifted so that people were generally less likely to respond that an item was a target at LP. Throughout the literature, the evidence shows that under LP conditions there is little change in *d'*, however, there is a consistent and robust shift in *c* (Wolfe et al., 2007, Van Wert, Horowitz & Wolfe, 2009, Wolfe & Van Wert, 2010, Russell & Kunar, 2012), resulting in a more conservative response bias. Please note that the majority of LP studies which have measured sensitivity have used *d'*. However, given the nature of prevalence research it has been suggested that instead it is better to calculate sensitivity using non-parametric measures (Wolfe & Van Wert, 2010, Godwin et al., 2010, Godwin et al., 2015). Accordingly, we calculate sensitivity using *A'* (see also Godwin et al., 2015 for an earlier use of this measure).

Understanding the mechanism underlying the LP Effect can help us understand how people search low prevalence displays in real world tasks - such as in mammography. However, in many real-world tasks there is often more than one person searching the display. For example, when searching for a cancer in mammograms in Australia, the UK and some other European countries, the recommended practice is that two readers search the same display (a process known as double reading, e.g., Perry et al., 2008, Wilson et al., 2011). The question of whether

double reading improves the efficiency of LP Search is an important one. If the benefit in double reading is not substantial, it may be that there are better ways of performing this type of search. For example, some countries use Computer Aided Detection (CAD) to help readers find cancers in mammogram reading (e.g., Bennett et al., 2006).

Previous research on the efficacy of double reading has provided mixed findings. For example, Taylor and Potts (2008) conducted a meta-analysis to investigate clinical double reading procedures and found that double reading leads to an increase in cancer detection rate and a decrease in recall rate (or false alarms, see also Taylor-Phillips et al., 2018). However, other studies which examined digital mammography concluded that there was no significant difference in either the rate of cancers detected or the rate of false positives between double reading and single reading conditions (Posso et al., 2017 see also Houssami et al., 2014, Sato et al., 2014 and Posso et al., 2016 who found limited benefits of using a double reading procedure). Please note that these studies examined double reading rates in a clinical setting. Therefore they were unable to determine the proportion of *miss errors* (as by definition radiologists would be unaware of cancers they have failed to detect). Miss errors typically can only be determined in the clinical setting retrospectively if a cancer is picked up at the next cancer screening or if the woman becomes symptomatic in the meantime. Instead, in order to determine miss errors we can run experiments in the laboratory, under conditions where we know that a cancer was present but participants failed to detect it.

Although there have been numerous lab-based studies that have investigated the LP Effect, the majority have used single reading procedures (i.e. only one person views the display). The one exception (that we know of) was conducted by Wolfe et al. (2007) who investigated the LP Effect using x-ray luggage images. In their experiments they had two participants search the

same display for a LP weapon and found that this led to a *small* reduction in miss errors. However, these experiments did not examine the contribution of motor-errors which could have inflated the magnitude of the observed LP Effect and thereby affected any reduction in miss errors found when double reading (Fleck & Mitroff, 2007). In the real-world motor errors would not be an issue as operators would have the opportunity to change their mind if they realised their mistake. Therefore, it is important to establish whether double reading can reduce the rate of miss errors caused by perceptual and/or attentional failures, rather than motor errors, per se. We investigate this here.

From a purely statistical perspective, having two readers search the same display should improve search performance given that there are two 'chances' of detecting the target. That is, having two people search a display will increase the probability that a target will be found. However, theoretically there are a number of different ways that having two readers could improve LP search (see Figure 1). First, it has been proposed that having two readers could improve sensitivity (Taylor & Potts, 2008, Gandomkar et al., 2018). In this case the combined reader potential would be better able to distinguish between signal and noise, or in terms of this search task better able to determine that a cancer is present and not just background breast tissue. We call this the *improved sensitivity* hypothesis. There is reason for this to be the case if we consider work from Mello-Thoms (2008) showing that individual readers in mammography often use different search strategies. Therefore, having two search strategies applied to the same display, might lead to an improvement in readers' ability to detect targets from non-targets (i.e., where one strategy fails to find the target, another one might locate it, thereby increasing overall sensitivity). In terms of signal detection an improvement in sensitivity would lead to an increase in *A'*. Second, having two readers could lead to a change in response bias *(c)*. Under LP conditions, response biases typically become more conservative

so people are less willing to say that a target is present. However, having two people read the display (especially if they are reading the display independently) might make the combined response more liberal (e.g. if one person says the target is absent, and the other says it is present, then in the clinical field the combined response would default to 'a cancer may be present' and go to the next level of screening, e.g., arbitration). We call this the *reduced response bias* hypothesis. Crucially, if this were the case then having two readers might mitigate the extreme change in response bias typically witnessed in LP search, which in turn would reduce the proportion of miss responses.

The work described in the current paper examines the benefit of having two people search a display for a rare target in comparison to only having one observer. We used a laboratory approach to enable us to compare *miss error rates*, along with false alarms across single reading and double reading conditions. Furthermore, in all experiments we gave participants an opportunity to self-correct their motor errors and had observers search a mammogram for a mass (using the search images developed by Kunar et al., 2017). This search task has the benefit of using search images from a real-world applied search task where the target is rare (e.g. within radiology) that are manipulated so that participants with non-medical backgrounds can perform the task[1]. In Experiment 1a, pairs of observers simultaneously searched the same display for a LP target on the same computer (the 'double reading' condition). The results were compared to a condition in which the same participants performed the search task on their own in different rooms (the 'single reading' condition). Experiment 1b was similar except that paired readers searched the same display independently (e.g. in different rooms). Together these experiments allowed us to assess any influence of spatial separation of the joint readers. Across this

[1] Participants were still given training in identifying the cancer prior to the experiment. However, they were not required to have the years of medical training radiologists needed for clinical reading. This allowed us to easily recruit and train more readers, thereby giving the experiments more power than in a clinical setting where it is often more difficult to recruit readers, due to their busy work schedule.

experiment we also examined whether any improvement in double reading occurred due to improved sensitivity or reduced response bias.

In Experiment 2, we again examined the performance of paired readers who searched the same display independently (e.g. in different rooms). However, here we used paired readers to assess whether there were vigilance decrements on the LP Effect. Given that LP tasks are often lengthy, it is possible that participants may suffer from vigilance 'drop-offs' leading to targets that appear at the end of the reading period being missed more often (e.g. See et al., 1995, Verster et al., 2013, Wiggins, 2011). Vigilance decrements, where attentional performance decreases with increasing time on task, have been observed in a wide variety of tasks, such as airport baggage screening and x-ray screening as well as in laboratory experiments (e.g. Basner et al., 2008, Taylor-Phillips et al., 2015, See et al., 1995). However, it is difficult to test vigilance decrements in LP search given the rarity of the target: as the proportion of targets in LP search is already very low (and hence very sparse) there are not enough target present trials for sensible analysis if the data are separated across time[2].

Taylor-Phillips et al. (2016) investigated whether a vigilance decrement occurred in clinical mammography by testing whether a change in reading order across readers in a double reading situation would affect cancer detection rates. In their Randomised Clinical Trial (RCT) they had a 'standard' condition where readers read a batch of mammograms in the same order (this replicates current double reading practice in the UK). They hypothesised that if a vigilance deficit occurred in this task then there would be a drop in performance in cancer detection rate

[2] Typical vigilance tests measure the proportion of targets that are missed as time on task increases, with the 'event rate' or number of targets presented per minute ranging from around 5 – 30 (see See et al., 1995 for a review). However, in LP search by definition the target only appears rarely (with an average 'event rate' of approximately 1 target every 6 – 12 minutes). This means that if the data were split into 30 minute segments, for example, there would only be approximately 2- 5 target present trials per time segment, which would be too few to analyse.

as time on task increased: as both readers saw the images in the same order this would mean that across the reader pair the same images would be affected by the same vigilance decline. This has the potential issue that women who have their mammograms read later in the session could be compromised in healthcare compared to those who have their mammograms read earlier in the session. Taylor-Phillips et al. (2016) compared this standard condition with an alternative in which paired readers saw the same mammograms, but the order was reversed for one of the readers. This would mean that any vigilance drop affecting the mammograms at the end of the reading batch in Reader 1, would be countered by the reversed viewing order for Reader 2, who saw these images first and therefore had not yet developed a vigilance decrement. Combining the participants' data, therefore, should offset any vigilance decrement to help cancer detection. However, Taylor-Phillips et al. (2016) found that, although time on task reduced the recall rate of individual readers (i.e. the number of women who were recalled for further tests), when the data were combined across readers, reading order did not affect the rates of breast cancers that were *detected*.

Please note that given the nature of RCTs there were several limitations that could have affected the outcome of this study. For example, Taylor-Phillips et al. (2016) noted that as data was collected over a large range of clinical reading centres they were unable to control for factors such as how many other mammograms had previously been read before the trial or the number of breaks that readers had been given: all of which could affect vigilance decrements. Furthermore, readers were only asked to read batches of 40 cases, which could mean that the reading session was too short to fully measure any vigilance deficit (reading time was hypothesised to take around 20 minutes, whereas vigilance drops can typically take 25- 35 minutes to be established, See et al., 1995). Lastly, as their study involved a clinical trial they were only able to measure the number of cancers detected in their study and were unable to

measure the effect of reading order on *miss errors* for cancers. Experiment 2, in this paper, replicated the work of Taylor-Phillips et al. (2016) task using a laboratory-based mammography task in which the proportion of miss errors could be determined. Furthermore, as our experiments were laboratory based they could be highly controlled in terms of number of breaks across participants and the number of mammograms that participants had read that day (i.e. none so that they entered the experiment non-fatigued). We also increased the reading batch from 40 cases to 1000 to allow sufficient time for any vigilance deficit to be observed (reading 1000 mammograms took approximately 2 hours to complete). Experiment 2 was similar to Experiment 1b except that half of the participant-pairs read each mammogram in the same order. The other participant pairs were presented with the mammograms in the opposite order (one participant read the mammogram batch forward, the other read it in reverse). We also changed the search task so that participants had to search for one mass taken from a range of potential targets to make the task more like that of a clinical setting (where a cancerous mass could take on a number of forms).

Experiment 1

Experiment 1 compared single reading to double reading procedures in LP search. In Experiment 1a, participant pairs searched the same display together in the same room. In Experiment 1b, participant pairs searched the same display in different rooms. The results of Experiment 1a and 1b were then combined. We predicted there would be a two-reader benefit with fewer miss errors in the double reading conditions compared to when participants searched the display alone (in the single reading conditions). The reason for this two-reader benefit was also examined by determining if it occurred due to an improved sensitivity or a reduced response bias account.

Method

Participants:

Twenty-four participants (M = 21.8 years, 10 female) took part in Experiment 1a and twenty-four participants (M = 21.3 years, 17 female) took part in Experiment 1b. In all experiments participants were recruited from the University of Warwick participant pool, had no prior training in reading mammograms and were paid for their time. In Experiment 1a participants were recruited in pairs. All participants had normal or corrected-to-normal vision. Ethical approval for all studies was granted by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick. Participant numbers were determined in advance based on previous research (e.g. Wolfe et al.,2007; Kunar et al., 2017). Prior to data collection a power analysis (F-tests effect size = 0.4, alpha = 0.05) showed that the minimum number of participants needed to achieve a power of 0.8 for each experiment was 16. Therefore, we would expect that testing 24 participants for each of the experiments would provide ample power to detect significant effects, if present.

Stimuli and Procedure:

The experiment was programmed using Blitz3D and presented on a PC. The mammogram images were taken from the selection of 'normal' mammograms (those not containing a cancer) of the Digital Database for Screening Mammography (DDSM) database (Heath et al., 2001, 1998). All images were selected from the database at random. Each LP condition contained 1000 images (2% target prevalence) and each HP condition contained 80 images (50% target prevalence). For each LP condition, 980 of these images were selected to act as target absent

trials for each HP condition 40 of these images were selected to act as target absent trials. Images were presented in the centre of the display and subtended approximately 11 degrees by 19 degrees at a viewing distance of 57 cm (although the individual size of each image varied because they were real mammograms)[3]. For target present trials an image of a cancerous mass was selected at random from one of the cancer cases on the DDSM and transposed onto the remaining mammogram images using imaging editing software[4]. The cancer could appear on any area of the breast tissue again chosen at random (mimicking conditions in a clinical setting), provided that it was clearly distinguishable once fixated (see Figure 2 for examples). All mammogram images were created offline.

Figure 2 about here

In Experiment 1a participants completed four conditions: a Single Reading Low Prevalence condition, a Single Reading High Prevalence condition, a Double Reading Low Prevalence condition and a Double Reading High Prevalence condition. In the Single Reading conditions participants completed each block as an individual (e.g. on different computers in different rooms)[5]. For both Low and High Prevalence conditions, a blank screen appeared for 500ms and was followed by a central grey fixation dot for 500ms. Following this one of the mammogram images was presented and remained on the screen until response (in this condition, presentation order of the images was randomised across participants). Participants

[3] Please note that some of the images from the DDSM contained dates and/or artefacts on the background of the image similar to images seen by radiologists in clinical mammography. However, as the dates/artefacts only appeared on the background of the image they did not affect the actual search task.
[4] In Experiments 1a and 1b only one example of a cancerous mass was used as a target. However, in Experiment 2 participants were asked to search one of a possible range of masses.
[5] Participants did not need to necessarily complete this condition at the same time as each other.

indicated whether the cancer was 'present' or 'absent' by pressing either the 'm' or the 'z' key respectively. Participants were instructed to respond as quickly but as accurately as possible and were informed of the prevalence rates of the target prior to the condition starting. If no response was made within 10 seconds, the trial 'timed-out' and the next trial started automatically. Following a response or 'time-out', a blank screen was again displayed before the next fixation dot and trial.

The Double Reading condition was similar, however participants completed the task in pairs. Prior to the blocks starting the pair agreed which participant was to press the response keys. Both participants sat in the same room and simultaneously viewed the same mammogram image on the same computer. They then had to verbally agree on whether the cancer was there or not before making a response. As in the Single Reading conditions, participant dyads completed both a HP and a LP condition.

Experiment 1b was the same as Experiment 1a, except that participants were tested individually in the Double Reading condition. That is participants completed the experiment in different rooms and were each responsible for their own responses[6]. The data were then paired, into participant-dyads after the experiment to give a joint response. This meant that for Experiment 1b, in target present trials, if one participant missed the target but the other correctly found it this was coded as a hit. If both participants failed to find the target this was coded as a miss. For target absent trials, if both participants identified the target as being 'absent' it was coded as a correct rejection, however, if either or both participants responded that the target was present it was counted as a false alarm. Participants were informed that in the Single Reading condition they were the only people to view the mammograms, whereas in the Double Reading

[6] Participants were not necessarily tested at the same time.

condition they were told that another participant would be viewing the images and that their data would be combined. Participants within a dyad were presented the images in the same order as each other (even though they were shown the images in different rooms), however across participant pairs the order of image presentation was random.

For both Experiments 1a and 1b, following Fleck and Mitroff (2007), participants had the option of correcting their responses in all conditions. If the participants recognized that they had made an error, they were able to correct it on the following trial, by pressing the 'Escape' key during any time of the next trial (see Fleck & Mitroff, 1997, Van Wert, et al., 2009, Kunar, et al., 2010, Kunar et al., 2017, Russell & Kunar, 2012, Rich et al., 2008, for similar methodologies). This would log in the data file that the participant had noticed their mistake so that motor errors could be calculated. They then proceeded with the current trial as normal, responding with an 'm' or 'z' key if the target was present or absent, respectively. No feedback was given after any response or correction was made and the experiments did not contain a reward mechanism in terms of point scoring for correct/incorrect responses.

For each of the high prevalence conditions (for both experiments) there were 80 trials with a 50% prevalence rate (40 present and 40 absent). For each of the low prevalence conditions, in which the target was present 2% of the time, there were 1000 trials (20 present and 980 absent). To familiarise themselves with the stimuli, participants were shown examples of the mammogram images and cancers prior to each of the experiments. They were also given a short practice block before each experimental block. During this practice block the experimenter ensured that participants were able to recognise the cancer, when present. If any of the participants had difficulties identifying the cancer they were shown more examples and could repeat the practice condition until both the participant and experimenter were confident that

they were able to identify the cancer. However, all the participants responded correctly in the first practice session and none were asked to repeat it. RTs, self-corrections and error rates were recorded. In the Low Prevalence blocks breaks occurred automatically every 200 trials, after which participants continued with the experiment when they were ready. Given the length of each experiment, the respective Single Reading and Double Reading conditions took place over two different sessions, each lasting approximately 2 hours. Within each experiment the presentation orders of the prevalence rates (HP versus LP) and reading condition (Single versus Double) were counterbalanced across participants.

Results

Due to a technical error 0.3 % of the data were corrupted and had to be removed as parts of the file were unreadable. This affected parts of the data files from three participants in Experiment 1b. The pattern of data remained unchanged if these participants were excluded from analysis. Initial errors and self-corrected error rates for all experiments can be found in Table 1. Consistent with Fleck and Mitroff (2007), for the single reading conditions of Experiment 1a, the ability to self-correct incorrect motor responses resulted in a significant reduction in miss errors for both the HP, $t(23) = 4.31$, $p < .001$, $d = 0.88$ and LP conditions, $t(23) = 4.52$, $p < .001$, $d = 0.92$. In the double reading conditions, miss errors were reduced after self-correction for the LP condition, $t(11) = 3.18$, $p = .01$, $d = 0.92$ but not for the HP one, $t(11) = 1.39$, $p = .19$, $d = 0.40$. For the single reading conditions of Experiment 1b, miss errors were reduced after self-correction for both HP, $t(23) = 2.77$, $p = .01$, $d = 0.57$ and LP, $t(23) = 4.49$, $p < .001$, $d = 0.92$. Comparisons could not be conducted for the double reading condition of Experiment 1b because the self-corrected miss error rates were zero in both HP and LP conditions. As we are primarily interested in cognitive rather than motor response errors, we focus our analysis

on the self-corrected data throughout the paper. For subsequent analyses, miss errors and false alarms (both for Experiment 1 and Experiment 2) were arcsine transformed prior to analysis to compensate for unequal variances present in binomial data (Hogg & Craig, 1995, see also Wolfe et al., 2007, Rich et al., 2008, Russell & Kunar, 2012, Kunar et al., 2010). Values reported and plotted in the figures are the back-transformed means. Miss errors and false alarm rates are shown in Figure 3 and mean correct RTs are shown in Table 2.

There are a number of possible statistics that we could conduct, however, for the purpose of this paper we concentrate our analyses on those that relate to our hypotheses. As we are also interested in how Single and Double reading affect LP search, we report planned t-tests comparing LP with HP conditions for miss errors, false alarms, *A'* and *c* for all experiments (see also, Wolfe et al., 2005, Wolfe et al., 2007, Fleck & Mitroff, 2007, Kunar et al., 2010, Kunar et al. 2017, Russell & Kunar, 2012, Rich et al., 2008, for similar analyses). In addition to frequentist statistics we report Bayes Factors analyses for these comparisons (calculated with a Cauchy prior width of 0.707 using JASP version 0.9.2)[7]. Bayesian analyses are presented alongside frequentist statistics as they have the advantage of being able to evaluate evidence in support of the null hypothesis (Wagenmakers et al., 2018a). We adopt the recommendations of Jeffreys (1961), in which a BF_{10} (which compares evidence of the alternative hypothesis to evidence for the null hypothesis) of 1 to 3 provides *anecdotal* evidence for the alternative, a BF_{10} of 3 to 10 provides *substantial* evidence for the alternative, a BF_{10} of 10 to 30 provides *strong* evidence for the alternative, a BF_{10} of 30 to 100 provides *very strong* evidence for the alternative and a BF_{10} of greater than 100 provides *decisive* evidence for the alternative. The

[7] Please note we only report Bayes statistics for the planned t-tests as Bayes factors for repeated measures ANOVAs still has its challenges and is an ongoing topic of research (Wagenmakers, et al. 2018b).

inverse of these numbers (BF_{01}) provide evidence in support the null hypothesis (Jarosz & Wiley, 2014).

Figure 3 and Tables 1 and 2 about here

Miss Errors

Examining the miss errors a 2 x 2 x 2 mixed-ANOVA with the within factor of Prevalence (High vs Low) and between factors of Condition (Double vs Single)[8] and Experiment (Experiment 1a: Same Room vs Experiment 1b: Different Room) revealed a main effect of Prevalence, $F(1, 68) = 28.75$, $p < .001$, $\eta p^2 = .30$, with more targets missed at LP than at HP and a main effect of Condition, $F(1, 68) = 16.67$, $p < .001$, $\eta p^2 = .20$, with more targets missed in the Single condition than in the Double reading condition. There was no main effect of Experiment, $F(1, 68) = 2.01$, $p = .16$, $\eta p^2 = .03$. The Prevalence x Condition interaction was significant, $F(1, 68) = 16.77$, $p < .001$, $\eta p^2 = .20$, with a larger LP Effect in the Single Reading than in the Double Reading condition. None of the other interactions were significant (all $Fs < 1$, $ps > .36$).

Planned t-tests showed that, for Experiment 1a, an LP Effect occurred in the Single Reading condition, $t(23) = -5.21$, $p < .001$, $d = 1.06$, with decisive evidence in support of the alternative, $BF_{10} = 864$, however there was no significant LP Effect in the Double Reading condition, $t(11) = -0.98$, $p = .35$, $d = 0.28$, with anecdotal evidence in support of the null, $BF_{10} = 0.43$. For Experiment 1b, an LP Effect occurred in the Single Reading condition, $t(23) = -5.5$, $p < .001$, $d = 1.12$, with decisive evidence in support of the alternative, $BF_{10} = 1540$. However, neither

[8] As data in the Double Reading conditions, by definition, were combined across pairs this led to an unequal number of data points between the Single and Double Reading condition. Therefore, for these and subsequent data analyses 'Condition' was treated as a between condition factor.

frequentist nor Bayesian t-tests could be conducted for the double reading condition of Experiment 1b because the miss errors were zero.

False Alarms

For the false alarms a 2 x 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Single vs Double) and Experiment (Experiment 1a: Same Room vs Experiment 1b: Different Room) showed there was no significant main effect of Prevalence, $F(1, 68) = 1.94$, $p = .17$, $\eta p^2 = .03$, nor of Condition, $F(1, 68) = 0.03$, $p = .87$ $\eta p^2 = .00$. There was a main effect of Experiment, $F(1, 68) = 5.83$, $p = .02$, $\eta p^2 = 0.08$, with more false alarms in Experiment 1b than in Experiment 1a. None of the interactions were significant (all $Fs < 3.26$, $ps > 0.07$). Overall, the false alarm rate was low (1.1%), consistent with previous work in which people searched for a clearly defined target (e.g. Wolfe et al., 2005, Kunar et al., 2010, Russell & Kunar, 2012, Kunar et al., 2017).

Planned t-tests across HP and LP conditions revealed that there was no significant effect of prevalence on false alarms in the single reading conditions for Experiments 1a, $t(23) = 1.33$, $p = .19$, $d = 0.27$ with anecdotal evidence in support of the null, $BF_{10} = 0.47$, or for Experiment 1b, $t(23) = 0.01$, $p = .99$, $d = 0.00$ with substantial evidence in support of the null, $BF_{10} = 0.22$. Neither was there an effect of prevalence on false alarms in the double reading conditions for Experiments 1a, $t(11) = 0.78$, $p = .45$, $d = .22$, with anecdotal evidence in support of the null, $BF_{10} = 0.37$, or for Experiment 1b, $t(11) = 1.07$, $p = .31$, $d = 0.31$, with anecdotal evidence in support of the null, $BF_{10} = 0.46$.

RTs

Mean correct RTs were analysed in a 2 x 2 x 2 mixed-ANOVA with the within factors of Prevalence (High vs Low) and Target Presence (Present vs Absent) and between factors of Condition (Single vs Double)[9] and Experiment (Experiment 1a: Same Room vs Experiment 1b: Different Room). There was a main effect of Prevalence, $F(1, 80) = 6.52$, $p = .01$, $\eta p^2 = .08$, with shorter RTs at HP than at LP. There was no significant main effect of Target Presence, $F(1, 80) = 0.07$, $p = .93$, $\eta p^2 = .00$, Condition, $F(1, 80) = 0.63$, $p = .43$, $\eta p^2 = .01$ nor of Experiment, $F(1, 80) = 2.54$, $p = .12$, $\eta p^2 = .03$. There was a significant interaction of Target Presence x Experiment, $F(1, 80) = 8.25$, $p = .01$, $\eta p^2 = .09$, in which the difference in RTs between Experiment 1a and Experiment 1b is larger for present trials than absent trials and a significant Prevalence x Target Presence x Experiment interaction, $F(1, 80) = 9.61$, $p = .003$, $\eta p^2 = .11$. Importantly, there was a significant Prevalence x Target Presence, $F(1, 80) = 160.31$, $p < .001$, $\eta p^2 = 0.67$, in which responses were faster at LP compared to HP when the target was absent, but slower at LP compared to HP when the target was present. This replicates previous work. None of the other interactions were significant (all $Fs < 2.04$, $ps > .15$).

Signal Detection Theory Analyses

In line with previous work, False Alarm and Hit rate data were used to calculate whether the LP Effect occurred due to a change in *A'* (reflecting a change in sensitivity) or *c* (reflecting a criterion shift) using Signal Detection Theory (SDT)[10]. Please note as mentioned in the Introduction we used *A'* to measure sensitivity rather than *d'* due to the nature of the data (see also Godwin et al., 2015 for similar analysis). Figure 4 shows the *A'* and *c* values.

[9] Please note that as participants in the Double condition of Experiment 1b completed the task separately we cannot meaningfully combine their data to produce an 'average' RT across the pair. Therefore, the double reading RTs here were used to see if there was a difference in RTs if participants believed another person was also performing the search task.

[10] False alarm or miss error rates of 0 and 1 were adjusted using the formulas $1/2n$ and $1-(1/2n)$, where n = the number of trials (Macmillan & Kaplan, 1985, see also Russell & Kunar, 2012, Wolfe et al., 2007 who used this procedure).

Figure 4 about here

Sensitivity (A')

Examining *A'* a 2 x 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Double vs Single) and Experiment (Experiment 1a: Same Room vs Experiment 1b: Different Room) revealed a main effect of Prevalence, $F(1, 68) = 22.87$, $p < .001$, $\eta p^2 = .25$, where *A'* was greater for HP than LP trials. There was a significant main effect of Condition, $F(1, 68) = 5.89$, $p = .02$, $\eta p^2 = .08$, where A' was greater for Double compared to Single reading conditions. However, there was no significant main effect of Experiment, $F(1, 68) = .04$, $p = .83$, $\eta p^2 = .0.001$. There was a significant Prevalence x Condition interaction, $F(1, 68) = 8.38$, $p = .005$, $\eta p^2 = .11$. None of the other interactions were significant, all Fs < 1, ps > .5.

Given that we are interested in whether an improved sensitivity or a reduced response bias account led to the double reading improvement, we separated the analysis into two separate ANOVAs for Experiments 1a and 1b. For Experiment 1a, a 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Single vs Double) revealed a significant main effect of Prevalence, $F(1, 34) = 13.13$, $p < .001$, $\eta p^2 = .28$, where *A'* was greater for HP than LP trials. There was also a significant effect of Condition, $F(1, 34) = 4.25$, $p = .047$, $\eta p^2 = .11$, where *A'* was greater for Double compared to Single reading. The Prevalence x Condition interaction was also significant, $F(1, 34) = 5.04$, $p = .03$, $\eta p^2 = .13$. Planned t-tests revealed that at LP, *A'* was greater in the Double reading compared to Single Reading condition, $t(34) = 2.18$, $p = .04$, $d = 0.77$, with anecdotal evidence in support of the

alternative, BF_{10} = 2.0. However there was no significant difference in *A'* between Double compared to Single reading at HP, $t(34) = 0.37$, $p = .71$, $d = 0.13$, with anecdotal evidence in support of the null, $BF_{10} = 0.35$.

For Experiment 1b, a 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Single vs Double) showed that there was a significant main effect of Prevalence, $F(1, 34) = 9.7$, $p = .004$, $\eta p^2 = .22$, where *A'* was greater for HP than LP trials. There was no main effect of Condition, $F(1, 34) = 1.97$, $p = .17$, $\eta p^2 = .06$. The Prevalence x Condition interaction was not significant, $F(1, 34) = 3.35$, $p = .08$, $\eta p^2 = .09$. Planned t-tests revealed that there was no difference in *A'* between the Double and Single Reading conditions either at LP or HP, $t(34) = 1.71$, $p = .10$, $d = 0.60$ and $t(34) = 0.20$, $p = .85$, $d = 0.07$, respectively, with similar evidence in support of the alternative and the null, BF_{10} = 1.005 for LP conditions and anecdotal evidence in support of the null for HP, $BF_{10} = 0.34$.

Criterion (c)

Examining *c*, a 2 x 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Double vs Single) and Experiment (Experiment 1a: Same Room vs Experiment 1b: Different Room) revealed a main effect of Prevalence, $F(1, 68) = 207.82$, $p < .001$, $\eta p^2 = .75$, showing that *c* was greater for LP trials than for HP. There was also a main effect of Experiment, $F(1, 68) = 10.56$, $p = .002$, $\eta p^2 = .13$, showing that *c* was greater for Experiment 1a than Experiment 1b. The main effect of Condition was not significant, $F(1, 68) = 3.69$, $p = .06$, $\eta p^2 = .05$. There was a significant interaction of Prevalence x Experiment, $F(1, 68) = 4.06$, $p = .048$,, $\eta p^2 = .06$, in which the difference in *c* across prevalence rates was greater in Experiment 1a than Experiment 1b, and a significant interaction between Condition x Experiment, $F(1, 68) = 6.55$, $p = .01$, $\eta p^2 = 0.09$, in which *c* was lower in the Double reading

compared to the Single reading condition in Experiment 1b but not in Experiment 1a. The Prevalence x Condition x Experiment interaction was also significant, $F(1, 68) = 5.62$, $p = .02$, $\eta p^2 = .08$.

As above, given that we are interested in whether an improved sensitivity or a reduced response bias account led to the double reading improvement, we separated the analysis into two ANOVAs for Experiments 1a and 1b. For Experiment 1a, a 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Single vs Double) revealed a main effect of Prevalence, $F(1, 34) = 127.20$, $p < .001$, $\eta p^2 = .79$, showing that *c* was greater for LP trials than for HP. There was no main effect of Condition $F(1, 34) = 0.332$, $p = 0.569$, $\eta p^2 = 0.01$, neither was the Prevalence x Condition interaction significant, $F(1, 34) = 0.09$, $p = .77$, $\eta p^2 = .003$. Planned t-tests revealed that there was no difference in *c* between Double and Single Reading for either LP or HP conditions $t(34) = 0.44$, $p = .67$, $d = 0.16$ and $t(34) = 0.44$, $p = .66$, $d = 0.16$, respectively, with anecdotal evidence in support of the null, $BF_{10} = 0.36$ for LP conditions and anecdotal evidence in support of the null for HP, $BF_{10} = 0.36$.

For Experiment 1b a 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and between factors of Condition (Single vs Double) revealed a main effect of Prevalence, $F(1, 34) = 81.91$, $p < .001$, $\eta p^2 = .71$, showing that *c* was greater for LP trials compared to HP. There was a main effect of Condition, $F(1, 34) = 7.25$, $p = .01$, $\eta p^2 = .18$, where responses were more conservative in the single condition than in the double condition. The Prevalence x Condition interaction was also significant, $F (1, 34) = 9.86$, $p = .003$, $\eta p^2 = .23$. Planned t-tests revealed that at LP, *c* was greater for Single reading compared to Double Reading conditions, $t(34) = 3.10$, $p = .004$, $d = 1.10$, with very strong evidence in support of the alternative, $BF_{10} =$

10.38. However there was no significant difference in *c* between Double and Single reading conditions at HP, $t(34) = 0.55$, $p = .59$, $d = 0.19$, with anecdotal evidence in support of the null, $BF_{10} = 0.38$.

Discussion

Experiment 1 investigated the benefit of having two people search an LP display either in the same room (Experiment 1a) or in a different room (Experiment 1b) compared to when they searched the display alone (in the single reading conditions). Furthermore, it was determined whether the two-reader benefit occurred due to an improved sensitivity account or a reduced response bias account for each of the different reading environments.

The data from the Single Reading conditions replicated the previous findings from the literature. False alarms were low, consistent with previous experiments in which participants searched for a well-defined target (Wolfe et al., 2005, Kunar et al., 2010, Russell & Kunar, 2012, Kunar et al., 2017). RTs also followed the typical LP pattern, in which RTs were faster at LP compared to HP when the target was absent but slower at LP compared to HP when the target was present. More importantly, the miss error data revealed an LP Effect whereby participants missed more targets at LP compared to HP.

Furthermore, having two people read the same mammograms led to fewer miss errors in the Double Reading conditions compared to the Single Reading conditions. In fact, after the data were combined across pairs in the Double Reading condition of Experiment 1b, the miss rates were zero. That is if one participant missed the target, the other participant in the reading pair detected it. There was a clear benefit from having a double reading methodology.

Interestingly, the reason for the double reading improvement differed across Experiments 1a and Experiment 1b. In Experiment 1a, *A'* was larger overall in the Double Reading condition compared to the Single Reading condition. When the data were separated across prevalence rates it showed that there was no benefit of double reading on *A'* at HP (please note that sensitivity was already high at HP), however, *A'* was greater in the double reading condition compared to the single reading condition at LP. This suggests that having two people read the same display in a shared environment led to an increase in sensitivity to find rare targets in line with an improved sensitivity account. We consider this finding further in the General Discussion. The data also showed that in both single and double reading conditions of Experiment 1a there was a change in response criteria: participants showed a more conservative response when target prevalence was low than when it was high. This replicates previous findings showing that under LP conditions participants were less willing to commit to a target present response. However, the difference in response bias across single and double reading conditions was not significant.

In contrast, the results of Experiment 1b revealed little difference in *A'* across single or double reading conditions. Having two people search the same display independently in *different rooms* did not improve sensitivity. Instead, the improvement in Experiment 1b was driven by how double reading affected response bias. Figure 4b reveals that, participants showed a more liberal response bias at LP in the Double compared with the Single reading condition. The results showed that when two people read the same display but in different rooms, although there was a shift in response bias to a more conservative response under LP conditions, this bias was less severe than in the single reading condition (when participants read the mammograms individually). The data are consistent with a reduced response bias account.

Having two people respond independently led to a 'second chance' procedure, where if one person missed the target then the other could have detected it. In this way combined responses were more likely to indicate that a target item was present, leading to fewer misses.

Please note that the sensitivity data in this experiment showed a different overall pattern than is typically found in the literature. Previous research has shown that sensitivity, as measured by *d'* does not change across prevalence rates (e.g., Wolfe et al, 2007, Russell & Kunar, 2012). However, in both Experiments 1a and 1b, sensitivity as measured by *A'* was greater overall for HP trials compared to LP. We discuss this further in the General Discussion.

The findings from Experiment 1 showed improved target detection rates with double reading compared to single reading. In Experiment 2 we investigated whether there was a further benefit to double reading if pairs of participants were given the mammogram images in different reading orders. Taylor-Phillips et al. (2016) examined this in a clinical setting to see whether any potential vigilance 'drop-off' (the detection of fewer targets with an increase of time-on-task), was removed by reversing the order the images were read between readers[11]. However, given that this was in the clinical setting they were unable to observe the proportion of cancers that were *missed*. Furthermore, Taylor-Phillips et al. (2016) also reported several limitations of their study in terms of being unable to control conditions across readers (e.g. the number of breaks that radiologists had, how many mammograms radiologists had read in a previous session – both of which could have affected the readers' vigilance). Therefore, in Experiment 2 we replicated the work by Taylor-Phillips et al. (2016) using our highly controlled laboratory task, allowing us to measure how miss errors were affected by having

[11] That is, if two readers read the same batch of mammograms in the same order they could both experience vigilance problems at the end of the session, which would affect the same mammogram images in both instances. Whereas, if the order of mammograms was different (for example, reversed) any vigilance decrement across reading pairs would be cancelled out.

participants view the images in the same order compared to in a different order (e.g. when the order was reversed for one of the pair).

Experiment 2 also increased the complexity of search by introducing multiple masses to search for. Instead of asking participants to search for one type of cancer, participants were asked to search for a range of different masses, including non-cancerous, benign masses. Please note that although in Experiment 2 there were a range of masses that could potentially act as the target, similar to Experiment 1, there was only ever one target that was present on target present displays. This had two advantages: first, it mimicked mammogram reading in a clinical setting more closely (where multiple types of masses - both cancerous and benign can appear) and second, as found by Kunar et al., (2007) having multiple targets to search for makes the search task more difficult and again better mimics search in real world mammography (see also Godwin, Menneer, Donnelly, & Cave, 2010, Menneer, Barrett, Phillips, Donnelly & Cave, 2007; Menneer, Cave, & Donnelly, 2009; Menneer, Donnelly, Godwin, & Cave, 2010, Kunar & Watson, 2011, Kunar & Watson, 2014). One of the potential reasons for the complete elimination of miss errors in Experiment 1b (in the double reading condition) may have been that the target was relatively easy to find[12]. By introducing a range of masses to search for in Experiment 2, we can examine the impact of having double readers when search becomes more difficult.

Experiment 2

[12] Although please note that even with a target that was relatively easy to find it was still missed more often at LP.

Taylor-Phillips et al. (2016) found that in a RCT, reading order had no effect of cancer detection. However, they were unable to determine how, or if, miss errors were affected by this procedure. Furthermore they had little control of reading conditions across clinical screening centres. Experiment 2 investigated whether reversing the reading order within pairs of readers would lead to a reduction in miss errors, using a highly controlled laboratory environment. Participants searched displays in participant pairs however, half of the participants searched the displays in the same order, the other half searched the displays in the reverse order. It was predicted that the results would replicate those of Taylor-Phillips et al. (2016) to show little effect of reading order on search.

Method

Participants:

Twenty-four participants (M = 21.6 years, 17 female) took part in the experiment. All had normal or corrected-to-normal vision.

Stimuli and Procedure:

The experiment was the same as Experiment 1b, except for the following changes. As Experiment 2 investigated manipulations to the order of reading in Double Reader conditions we did not include a Single Reading condition (you cannot compare 'Same Order' or 'Reverse Order' conditions with only one reader). For the Double Reading condition, half the participants viewed the images in the same order as each other (the Same Order condition). The remaining participants viewed the images in reverse of each other (the Reverse Order

condition). That is if one of the participants saw the images in the order of 1…n, the other participant in the dyad saw the images in the order of n…1. For each dyad a new presentation order was randomly generated so that across dyads the order of presentation was varied. However, within the dyad pair the order of the image presentation was either the same or reversed. For all conditions, participants viewed the images in a different room to their partner.

The stimuli were also changed so that the mass could either be a benign mass or a cancerous mass (see Kunar et al., 2017, for examples). To create these displays, 80 'normal' mammograms (those not containing a cancer) were randomly selected from the DDSM (40 for HP trials and 40 for LP trials). These images were then digitally edited to include either a cancerous mass or a benign mass. Four masses were selected from the Cancer cases and four masses were selected from the Benign cases of the DDSM. Each mass was then transposed onto ten of the 'normal' mammogram images to create 40 cancerous mass mammograms (20 to be used in the HP condition and 20 to be used in the LP condition) and 40 benign mass mammograms (20 to be used in the HP condition and 20 to be used in the LP condition). Similar to Experiment 1, the masses could appear on any area of the breast tissue, as long as it was clearly distinguishable once fixated. Example images can be seen in Figure 5. All mammogram displays were created offline. Target absent trials were created in a similar manner to Experiment 1 by randomly selecting from the DDSM 40 'normal' mammograms for the HP condition and 960 'normal' mammograms for the LP condition.

Figure 5 about here

For the High Prevalence conditions, where a mass was present 50% of the time, there were 80 trials. Half of these trials were target present and contained either a cancerous or benign mass (20 trials with a cancerous mass and 20 trials with a benign mass). The other half of the displays were target absent displays. For the Low Prevalence conditions there were 1000 trials (20 trials with a cancerous mass, 20 trials with a benign mass and 960 absent target images). This meant that although there was a mass present 4% of the time, the overall probability of a cancerous image being encountered was 2%. To familiarise themselves with the stimuli, participants completed a session in which they were shown examples of the mammogram displays and instructed how to identify both benign and cancerous masses prior to the experiment. During this session, participants were shown example displays of mammogram images and asked to locate and identify each mass. Once the experimenter was confident that the participant could identify the masses, they completed a training test before the experiment proper. The training test confirmed participants' ability to recognise and classify a mass as either benign or cancerous (by pressing 'b' or 'c' on a computer keyboard) and included 24 examples of mammogram displays (12 containing a cancer and 12 containing a benign mass). Participants only continued onto the experiment once they had completed the training test and the experimenter had determined that they could correctly identify each mass.

The procedure was similar to that used in Experiment 1. Participants were asked to indicate whether a mass (either cancer or benign) was present or absent by pressing either the 'm' or the 'z' key respectively and given the option to self-correct by pressing the escape key on the following trial. However, if they pressed the present key they were then given a follow up question asking them to identify the mass as either cancerous or benign by pressing either the

'c' or the 'b' key respectively[13]. The presentation order of prevalence rates (HP versus LP) was counterbalanced across participants.

Results

Experiment 2 differs from Experiment 1, in that participants are asked to search for a range of masses, rather than just one mass. To determine the effect of double reading on conditions where participants need to search for multiple masses it is first important to show that the typical LP Effect occurs with these stimuli. As Experiment 2 did not contain a single reading condition we did this by analysing the non-combined data to show an LP Effect occurred (Table 3). From these data we can also determine the effect of LP on mass identification. Following this we then combined the data across paired readers to examine how reading order effects mass detection in double reading.

Replication of Prevalence Effect when looking for Multiple Targets

Miss Errors

Examining the miss errors of the non-combined data, a 2 x 2 within-subjects ANOVA factors of Prevalence (High vs Low) and Mass Type (Benign vs Cancer) revealed a significant main effect of Prevalence, $F(1, 23) = 12.34$, $p = .002$, $\eta p^2 = .35$, where more targets were missed at LP than at HP. There was also a main effect of Mass Type, $F(1, 23) = 23.05$, $p < .001$, $\eta p^2 = .50$, with more cancers missed than benign targets. However, the Prevalence x Mass Type interaction was not significant, $F(1, 23) = 1.62$, $p = .22$, $\eta p^2 = .07$.

[13] Please note there was no self-correction button for target identification as these responses would be unlikely to be influenced by speed-error trade-offs that effect LP detection responses (Fleck & Mitroff, 2007).

Planned t-tests showed that an LP Effect occurred when the target was benign, $t(23) = -2.45$, $p = .02$, $d = 0.50$, with anecdotal evidence in support of the alternative, $BF_{10} = 2.47$, and also when the target was a cancer, $t(23) = 3.33$, $p = .003$, $d = 0.68$, with strong evidence in support of the alternative, $BF_{10} = 13.86$.

False Alarms

A t-test on false alarms (HP vs LP)[14] rates showed there was no effect of Prevalence, $t(23) = 0.40$, $p = .69$, $d = 0.08$, with substantial evidence in support of the null, $BF_{10} = 0.23$. However, false alarms were higher in this experiment compared to Experiments 1a and 1b. This was confirmed when we compared single reading false alarm rates across experiments. A 2 x 2 ANOVA with within-subject factors of Prevalence (High vs Low) and between-subject factors of Experiment (Experiment 1a vs Experiment 2) revealed a significant main effect of Experiment, $F(1, 46) = 32.97$, $p < .001$, $\eta p^2 = 0.42$, where there were more false alarms in Experiment 2 than in Experiment 1a. There was no main effect of Prevalence, $F(1, 46) = 0.07$, $p = .79$, $\eta p^2 = .002$, nor a significant Prevalence x Experiment interaction, $F(1, 46) = 0.95$, $p = .34$, $\eta p^2 = .02$. Similarly, when comparing Experiment 2 and Experiment 1b there was a main effect of Experiment, $F(1, 46) = 23.08$, $p < .001$, $\eta p^2 = .56$, with more false alarms in Experiment 2 than in Experiment 1b. There was no main effect of Prevalence, $F(1, 46) = 0.15$, $p = .70$, $\eta p^2 = .003$, or a Prevalence x Experiment interaction $F(1, 46) = 0.14$, $p = .71$, $\eta p^2 = .003$.

RTs

[14] As the False Alarm data only used target absent trials then unlike the Miss Error data, False alarms could not be split up into Mass Type.

Given that target absent trials could not be meaningfully categorised according to 'Mass Type' (as there was by definition no mass present when the target was absent) RTs were analysed separately for target absent and target present trials (Table 2). Examining the RTs for target present trials a 2 x 2 ANOVA with the within factors of Prevalence (High vs Low) and between factors of Mass Type (Benign vs Cancer) revealed a main effect of Prevalence, $F(1, 23) = 12.37$, $p = .002$, $\eta p^2 = .35$ with shorter RTs at HP than at LP. There was also a main effect of Mass Type, $F(1, 23) = 22.94$, $p < .001$, $\eta p^2 = .50$, in which RTs for benign masses were shorter than those for cancers. The Prevalence x Mass Type interaction was not significant, $F(1, 23) = 0.24$, $p = .63$, $\eta p^2 = .01$.

Examining RTs for target absent trials a t-test revealed an effect of Prevalence in which RTs were faster in LP trials compared to HP, $t(23) = 2.07$, $p = .05$, $d = 0.42$, with anecdotal evidence in support of the alternative, $BF_{10} = 1.30$.

Sensitivity (A')

Examining *A'* a 2 x 2 ANOVA with within factors of Prevalence (High vs Low) and Mass Type (Benign vs Cancer) showed that there was no main effect of Prevalence, $F(1, 23) = 2.51$, $p = .13$, $\eta p^2 = .01$. There was a main effect of Mass Type, $F(1, 23) = 24.98$, $p < .001$, $\eta p^2 = 0.52$, where *A'* was greater for benign compared to cancerous masses. The Prevalence x Condition interaction was significant, $F(1, 23) = 4.65$, $p = .04$, $\eta p^2 = .17$. Planned t-tests showed that there was no effect of prevalence on *A'* for benign targets, $t(23) = 0.63$, $p = .53$, $d = 0.13$, with substantial evidence in support of the null, $BF_{10} = 0.26$. However, *A'* was greater at HP than LP for cancerous masses, $t(23) = 2.17$, $p = .04$, $d = 0.44$, with anecdotal evidence in support of the alternative, $BF_{10} = 1.54$.

Criterion (c)

Examining *c*, a 2 x 2 ANOVA with within factors of Prevalence (High vs Low) and between factors of Mass Type (Benign vs Cancer) revealed a main effect of Prevalence, $F(1, 23) = 6.86$, $p = .02$, $\eta p^2 = .23$, showing that responses were more conservative for LP trials compared to HP. There was a main effect of Mass Type, $F(1, 23) = 30.61$, $p < .001$, $\eta p^2 = 0.57$, where responses were more conservative for cancer targets than benign. The Prevalence x Condition interaction was not significant, $F(1, 23) = 2.64$, $p = .12$, $\eta p^2 = 0.10$. Planned t-tests showed that *c* was higher at LP than HP for both benign targets, $t(23) = 2.17$, $p = .04$, $d = 0.44$, with anecdotal evidence in support of the alternative, $BF_{10} = 1.55$, and for cancers, $t(23) = 2.87$, $p = .009$, $d = 0.59$, with substantial evidence in support of the alternative, $BF_{10} = 5.46$.

Mass Identification Errors

As participants were asked to identify each mass after they had detected it, in this experiment we can also measure how accurately they did this. Overall, participants incorrectly identified the mass 25.8% of the time. A 2 x 2 ANOVA with within factors of Prevalence (High vs Low) and Mass Type (Cancer vs Benign) revealed a main effect of Prevalence, $F(1, 23) = 10.13$, $p = .004$, $\eta p^2 = .31$, where participants were worse at identifying the mass at LP compared to HP (Identification Errors = 29.0% vs 22.6%, respectively). There was no main effect of Mass Type, $F(1, 23) = 0.60$, $p = .45$, $\eta p^2 = .03$. Neither was the Prevalence x Mass Type interaction significant, $F(1, 23) = 1.12$, $p = .30$, $\eta p^2 = .05$. Planned t-tests revealed that there was no significant effect of identification errors across prevalence for benign masses (29.6% vs 25.3%, respectively), $t(23) = 1.84$, $p = .08$, $d = 0.38$, with anecdotal evidence for the null, $BF_{10} = 1.10$. However, mass identification was significantly worse for cancers at LP than HP (28.3% vs 20.0% errors, respectively), $t(23) = 2.68$, $p = .01$, $d = 0.55$, with substantial evidence for the alternative, $BF_{10} = 3.81$.

The results from the non-combined data confirmed the presence of an LP Effect when participants were asked to search for a range of multiple targets. Furthermore, participants' ability to identify the mass was affected by target prevalence, so that people were worse at identifying cancers at LP. The data were then combined across participant pairs to allow us to examine how reading order affected performance when participants were double reading. Figure 6 shows the miss error and false alarm data and Figure 7 shows the *d'* and *c* values.

Figures 6 and 7 about here

Paired Reading: Same order vs Reverse order

Miss Errors

Examining the miss errors, a 2 x 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and Mass Type (Cancer vs Benign) and between factors of Reading Order (Same vs Reverse) revealed a main effect of Mass Type, ($F(1, 10) = 10.53$, $p = .009$, $\eta p^2 = .51$, where more cancerous targets were missed than benign masses. There was no main effect of Prevalence, $F(1, 10) = 4.43$, $p = .06$, $\eta p^2 = .31$ or Reading Order, $F(1, 10) = 0.97$, $p = .35$, $\eta p^2 = .09$. The Prevalence x Mass Type interaction was significant, $F(1, 10) = 9.20$, $p = .01$, $\eta p^2 = 0.48$, more masses were missed at LP compared to HP for cancerous targets but not benign. No other interactions, including those with Reading Order were significant (all Fs < 1, ps > 0.48). Planned t-tests revealed that there was no LP Effect in when the target was benign, $t(11) = 1$, $p = .34$, $d = 0.29$, with anecdotal evidence in support of the null, $BF_{10} = 0.44$. However, there was an LP Effect when the target was a cancer, $t(11) = 2.86$, $p = .02$, $d = 0.83$, with substantial evidence in support of the alternative, $BF_{10} = 4.15$.

False Alarms

For the false alarms, a 2 x 2 ANOVA with within factors of Prevalence (High vs Low) and between factors of Reading Order (Same vs Reverse) showed that there was no main effect of Prevalence, $F(1, 10) = 0.08$, $p = .79$, $\eta p^2 = .01$, or Reading Order, $F(1, 10) = 0.18$, $p = .68$, $\eta p^2 = .02$, and the Prevalence x Reading Order interaction was also non-significant, $F(1, 10) = 0.77$, $p = .40$, $\eta p^2 = .07$. A planned t-test revealed that there was no effect of prevalence on false alarms, $t(11) = 0.28$, $p = .78$, $d = 0.08$, with substantial evidence in support of the null, $BF_{10} = 0.30$.

Sensitivity (A')

Examining *A'* a 2 x 2 x 2 mixed-ANOVA with within-subjects factors of Prevalence (High vs Low) and Mass Type (Cancer vs Benign) and between-subjects factors of Reading Order (Same vs Reverse) revealed that there was no main effect of Prevalence, $F(1, 10) = 0.009$, $p = .93$, $\eta p^2 = 0.001$. There was, however, a main effect of Mass Type, $F(1, 10) = 15.76$, $p = .003$, $\eta p^2 = .61$, where *A'* was greater for benign masses compared to cancers. There was no significant effect of Reading Order on *A'*, $F(1, 10) = 0.33$, $p = .58$, $\eta p^2 = .03$. The Prevalence x Mass Type interaction was significant, $F(1, 10) = 12.68$, $p = .01$, $\eta p^2 = .56$. None of the other interactions were significant (all Fs < 1.58, ps > .23). Planned t-tests revealed that there was no significant difference in *A'* across prevalence for benign masses, $t(11) = 0.58$, $p = .57$, $d = 0.17$, with substantial evidence in support of the null, $BF_{10} = 0.33$, nor was there a difference in *A'* across prevalence for cancerous masses, $t(11) = 0.41$, $p = .69$, $d = 0.12$, with substantial evidence in support of the null, $BF_{10} = 0.31$.

Criterion (c)

Examining *c*, a 2 x 2 x 2 mixed-ANOVA with within factors of Prevalence (High vs Low) and Mass Type (Cancer vs Benign) and between factors of Reading Order (Same vs Reverse) revealed that there was no significant main effect of Prevalence, $F(1, 10) = 0.69$, $p = .43$. $\eta p^2 = .06$. There was a main effect of Mass Type, $F(1, 10) = 8.55$ $p = .02$, $\eta p^2 = .46$, where *c* was greater for cancers compared to benign masses. There was no effect of Reading Order, $F(1, 10) = 0.06$, $p = .81$, $\eta p^2 = .01$. The Prevalence x Mass Type interaction was significant, $F(1, 10) = 9.83$, $p = .01$, $\eta p^2 = .50$. None of the other interactions were significant (all Fs < 1.97, ps > .19). Planned t-tests revealed that there was no significant difference in *c* across prevalence for benign masses, $t(11) = 0.33$, $p = .75$, $d = 0.10$, with substantial evidence in support of the null, $BF_{10} = 0.30$, nor was there a difference in *c* for cancerous masses, $t(11) = 1.26$, $p = .23$, $d = 0.37$, with anecdotal evidence in support of the null, $BF_{10} = 0.55$.

Discussion

Experiment 2 examined whether, under controlled laboratory conditions, miss error rates could be reduced by having a pair of readers each see the mammograms in a different order rather than seeing them in the same order. The different order technique should reduce any performance decrements caused by vigilance deficits. However, the results showed that reading order did not affect miss errors: there was little difference in the proportion of masses missed between the same reading condition and the reverse reading condition. This extends the findings of Taylor-Phillips et al. (2016), who found little difference in *cancers detected* across reading order, however our results have been able to determine a similar effect with the proportion of masses missed in an experimentally controlled environment.

Examining error rates, an LP Effect occurred overall. This occurred in the single reading data (before it was combined) and also after the data were combined into paired readers, but only for trials when the target was a cancer. Although there was no significant difference in false alarms across prevalence rates, false alarms were higher overall than previous experiments in this paper. This fits with other work showing that increasing the number of possible targets leads to a greater number of errors (e.g., Kunar et al., 2017, Godwin, Menneer, Donnelly, & Cave, 2010, Menneer, Barrett, Phillips, Donnelly & Cave, 2007; Menneer, Cave, & Donnelly, 2009; Menneer, Donnelly, Godwin, & Cave, 2010, Kunar & Watson, 2011, Kunar & Watson, 2014).

In this experiment we also introduced different types of masses (cancer and benign) into the search task. The results showed that more cancers were missed than benign targets. Although this is of potential interest we do not wish to put much weight on it as it could have resulted from the type of examples we used. As our participants did not have medical training we made sure our benign and cancerous masses were perceptually distinct from each other so that they could be distinguished in a laboratory setting. For example, in our study the benign masses were more uniform in their texture and were less spiculated than cancers. This might have led these particular masses to be more easily detected in the search task – however, this result may not translate to real world mammography where the benign and cancerous masses will be more variable. Despite this, we can use the data to compare Mass Identification across prevalence rates as the same examples of masses were used in both HP and LP trials (the only difference being their relative prevalence rates). Interestingly, participants made more errors when identifying masses at LP than at HP. Not only does prevalence effect target detection – it can also effect identification.

General Discussion

Across two experiments we investigated the effects of double reading on Low Prevalence search within a simulated mammography task. Experiment 1 showed that having two readers perform the task led to fewer miss errors compared to when only one reader viewed the displays. Experiment 2 showed that there was little benefit of having readers view the images in different orders. All experiments had a self-correction option (Fleck & Mitroff, 2007). This allowed us to determine a more accurate measure of the proportion of *perceptual* miss errors without the inclusion of motor errors, which are easily rectified in the clinical field. The LP Effect is known to be robust and notoriously difficult to alleviate (e.g. Wolfe et al., 2007). Given our findings, however, we propose there is substantial improvement in LP tasks of having two, rather than one, reader.

Interestingly, the SDT data suggested that the reasons why double reading reduces miss errors differ across reading conditions. When two people searched the same display independently (i.e., in different rooms in Experiment 1b) the data point to a *reduced response bias* account where combined response pairs modulated the large shift in criteria typically observed under LP search. Having two people respond independently to create a combined response led to a more liberal response bias than that found in single reading conditions, which in turn led to fewer missed errors. Please note that this change in response bias only occurred when each participant was asked to search the display and produce their own individual response: when participants gave a joint response (Experiment 1a) no shift in response bias was observed between single and double reading conditions. Having independent responses meant that the final combined response was counted as a 'hit' even if only one person in the pair responded 'target present'. This is also true within clinical settings. For example, in cases where only one

reader detects a potential cancer the woman may still be recalled for further tests or the case is sent to arbitration, to be examined further by another reader or group of readers.

In contrast to the reduced response bias account found in Experiment 1b, the data from Experiment 1a showed that having a shared double reading environment led to an improvement in sensitivity. There could be a number of reasons for this. First, Brennan et al. (2008) suggested that peoples' ability to collaboratively search together improved if they were able to share gaze and vocal information. That is the knowledge of your partner's gaze and speech communication led to better and more efficient search than searching the display alone. In our experiments, participants were only able to communicate with each other in Experiment 1a, where they shared a room (in Experiment 1b although participants were aware other people were searching the display they were not able to communicate with them). This increase in communication may have led to an improvement in sensitivity (e.g. one person may suggest a suspicious area, which both participants then jointly agree to be a target).

Second, the physical presence of another person in the room may have led to improved task performance. This is known as the 'social facilitation' effect, where people are often better at performing a task when they are in the presence of another person compared to when they are performing the task unwatched (e.g., Bond & Titus, 1983). Chib et al. (2018) recently examined the social facilitation effect with the use of functional Magnetic Resonance Imaging (fMRI) and found that performing a task in the presence of another person increased activity in and between the dorsomedial prefrontal cortex (dmPFC) and the ventromedial prefrontal cortex (vmPFC) – areas that are thought to be involved in reward outcomes, motivation and social cognition (Liljeholm and O'Doherty, 2012). Chib et al. (2018) suggested that when participants completed a task in the presence of another person there were social motivational signals,

which led to people being better motivated to perform the task well compared to when they are not being watched. This social facilitation may explain why people in our experiments showed an increase in sensitivity when physically performing the task with another person. It will be up to future research to examine this further.

The results from Experiment 2 showed that participants made more errors when asked to *identify* the target at LP compared to HP. In this case, participants had to make a 2AFC response as to whether the mass was cancerous or benign. Across HP and LP conditions, the actual number of target present trials was identical, instead under LP, by definition participants saw the targets less frequently. It could be that our ability to perceptually categorise items is better if they appear frequently compared to when they only appear rarely. Studies have shown that developmentally our brains are wired to better perceive items that appear often compared to items that have infrequent exposure (e.g., Lewkowicz & Ghazanfar, 2009). Therefore, it may be that under HP trials as the masses appeared more frequently participants were better able to categorise them as benign or cancerous, leading to fewer identification errors. In fact, in the clinical field it has been found that examining mammograms in batches all at one time, rather than examining individual cases one at a time leads to better cancer screening (Burnside et al., 2005). Please note, that these results differ from Kunar et al. (2017) who found that although target identification errors were affected by Computer Aided Detection cues, in contrast to the work here, there was no effect of prevalence rates. Therefore, we treat this result cautiously and future research is needed to investigate this further.

Previous research has found that sensitivity as measured by *d'* does not differ between LP and HP conditions. However, the data from Single Reading conditions in Experiments 1a and 1b and the non-combined cancerous mass conditions in Experiment 2 showed that sensitivity as

measured by *A'* was greater at HP than LP. There are two potential reasons for this. First it might be that using the non-parametric measurement of *A'*, rather than *d'* provides different results in terms of sensitivity change across prevalence. In fact, if we analyse our data using *d'* as a measure we see no difference in sensitivity across prevalence rates in the Single Reading conditions[15]. However, this cannot account for all of the data in the field as although the majority of low prevalence SDT analysis have used *d',* Godwin et al. (2015) used *A'* to measure sensitivity and found little difference in *A'* across prevalence rates. Second it could be that our particular search task (e.g. search for cancerous masses in mammograms) led to changes in sensitivity across prevalence rates. This might be the case given that cancers are often difficult to detect depending on the density of the surrounding breast tissue. It will be up to future research to investigate this further. Please note that the drop in sensitivity we found at LP only occurred in the Single Reading and not Double reading conditions.

Our research shows a clear benefit of the double reading procedure. However, it is also important to note that there are financial and practical implications associated with employing two readers in a clinical setting. For example, Guerriero et al. (2011) stated that, in mammography, employing a double reading procedure can lead to increased demands on radiologists' time, which may be difficult to sustain with an increasing number of women who need screening. Furthermore, Posso et al. (2016) found that using a double-reading program in Spain was not financially effective. Given these limitations, it is also important to investigate other ways to help readers find LP targets. One alternative is to use Computer Aided Detection (CAD), whereby computer algorithms are used to flag up 'suspicious' areas to the reader for further attention. However, the benefits of CAD are still being investigated and remain

[15] Please note the pattern across Double and Single Reading conditions in sensitivity was the same with both D' and A' measurements.

controversial (Fenton et al., 2007; Philpotts, 2009). For example, in a large-scale meta-analysis, using data from eight studies, Bennett et al. (2006) found that the benefits or costs of CAD were inconclusive. Guerriero et al. (2011) also found that the implementation of CAD, at least in the UK, was not financially cost-effective (given the associated extra costs of equipment, training and maintenance). Furthermore, work from our lab has shown that CAD can lead to an *increase* in miss errors as people over-rely on the technology which becomes problematic when CAD cues fail (Kunar et al., 2017). With technological advancements it may be that CAD systems will overcome these issues in the future so that their detection ability out-performs that of a trained human reader. However, for now techniques such as double reading are an effective way to reduce miss error rates in LP search.

Taylor-Phillips et al. (2016) found that there was no significant advantage to having double readers see the images in the same or in opposite orders on cancer detection. Here we were able to replicate and extend these findings by also examining reading order on *miss errors* in a more tightly controlled environment, which could not be measured in a clinical setting. Our results complement those from the randomised clinical setting. This is assuring given the concern that laboratory research could yield different results to real world search tasks (e.g., Gur et al., 2003). Other work has also shown that results found in the laboratory can also be found in a clinical setting. Evans et al. (2013) examined whether radiologists exhibit the LP Effect by embedding a series of mammograms showing a cancer into the normal breast cancer screening service in a hospital setting (to create a known prevalence rate of ~ 1%). Radiologists viewed cases over a period of nine months, thereby reading a large number of scans across this time period. The results showed that even in this situation an LP Effect occurred where radiologists missed 30% of these cancer cases. This is an important finding given that there are a number of methodological differences between laboratory experiments like ours and procedures used

in clinical settings. For example in our experiments participants read a greater number of mammograms in one sitting than would typically be read in a breast screening centre. However, despite these differences in methodologies, given the similarity between Evans et al. (2013) findings in a clinical setting and those found in the laboratory, we believe the results found in this paper have relevance to the clinical field.

At present, mammogram reading procedures vary world-wide. The National Health Service in the UK for example has made double reading of digital mammography mandatory (Wilson et al., 2011), whereas in Australia, although not mandatory, double reading is considered preferable (The Royal Australian and New Zealand College of Radiologists, 2014). Despite this, other countries do not require double reading (for example, in the USA the decision rests with individual centres, Taylor-Phillips, 2016). Given our results we suggest that best practice would be to use a double reading procedure, for tasks like mammography, universally.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. Journal of Experimental Psychology: General, 120, 3–19.

Bennett RL, Blanks RG, Moss SM. (2006) Does the accuracy of single reading with CAD (computer-aided detection) compare with that of double reading?: A review of the literature. Clin Radiol.;61(12):1023-8.

Burnside ES, Park JM, Fine JP, Sisney GA. (2005) The use of batch reading to improve the performance of screening mammography. *AJR AmJ Roentgenol*, 185(3):790-796.

Chelazzi L, Perlato A, Santandrea E, Della Libera C. (2013) Rewards teach visual selective attention. Vision res. 85:58–72

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. *PloS one, 8(5),* e64366.

Fenton JJ , Taplin SH , Carney PA , et al (2007) . Influence of computer-aided detection on performance of screening mammography . *N Engl J Med; 356* (14): 1399 – 1409 .

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science , 18* (11), 943-947.

Gandomkar, Z., Tay, K., Brennan, P. C., Kozuch, E., & Mello-Thoms, C. R. (2018). Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task? *Medical Physics*, *45*, 4844–4956.

Godwin, H.J., Menneer, T., Donnelly, N. and Cave, K.R. (2010) Dual-target search for high and low prevalence x-ray threat targets. *Visual Cognition*, 18, (10), 1439-1463.

Godwin, H. J., Menneer, T., Cave, K. R., Thaibsyah, M., & Donnelly, N. (2015). The effects of increasing target prevalence on information-processing during visual search. *Psychonomic Bulletin & Review*. 22:469–475

Green, D. M., & Swets, J. A. (1967). Signal detection theory and psychophysics. New York: John Wiley and Sons.

Guerriero, C , Gillan, MGC, Cairns, J, Wallis, MG, and Gilbert, FJ (2011) Is computer aided detection (CAD) cost effective in screening mammography? A model based on the CADET II study. *BMC Health Serv Res.; 11: 11*. Published online 2011 January 17. doi: 10.1186/1472-6963-11-11

Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, et al. (2003) Prevalence effect in a laboratory environment. *Radiology 228* 10–14.

Heath, M., Bowyer, K., Kopans, D., Moore, R. and Kegelmeyer, W.P. (2001) *Proceedings of the Fifth International Workshop on Digital Mammography*, M.J. Yaffe, ed., 212-218, Medical Physics Publishing, ISBN 1-930524-00-5.

Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, W.P., Moore, R., Chang, K. and MunishKumaran, S. (1998) *Digital Mammography*, 457-460, Kluwer Academic Publishers, Proceedings of the Fourth International Workshop on Digital Mammography.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (5th ed.). Englewood Cliffs, NJ: Prentice Hall International.

Houssami N, Macaskill P, Bernardi D, et al. (2014). Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading– evidence to guide future screening strategies. *European Journal of Cancer. ;50(10):*17991807.

Kunar, M. A., Ariyabandu, S. & Jami, Z. (2016). The Down Side Of Choice: Having A Choice Benefits Enjoyment But At A Cost To Efficiency and Time In Visual Search. *Attention, Perception and Psychophysics, 78,* 736-741.

Kunar, M.A., Rich, A.N. & Wolfe, J.M. (2010). Spatial and temporal separation fails to counteract the effects of low prevalence in visual search. *Visual Cognition, 18,* 881-897.

Kunar, M. A., Watson, D.G., Taylor-Phillips, S. & Wolska, J. (2017). Low Prevalence Search for Cancers in Mammograms: Evidence using Laboratory Experiments and Computer Aided Detection. *Journal of Experimental Psychology: Applied, 23,* 369-385.

Kunar, M. A., Watson, D.G., Tsetsos, K. & Chater, N. (2017). The Influence of Attention on Value Integration. *Attention, Perception & Psychophysics, 79,* 1615-1627.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. Trends in Cognitive Sciences, 13(11), 470-478.

Macmillan, N. A. & Creelman, C. D. (2005). Detection Theory: A User's Guide, 2nd Edition, Cambridge University Press.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of grou*p* data: estimating sensitivity from average hit and false-alarm rates. Psychological Bulletin, 98, 185-199.

Mello-Thoms C: How Much Agreement is There in the Visual Search Strategy of Experts Reading Mammograms? SPIE, San Diego, 2008

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly, N., & Cave, K. R. (2007). Costs in searching for two targets: Dividing search across target types could improve airport security screening. *Applied Cognitive Psychology*, *21*, 915-932.

Menneer, T., Cave, K. R., & Donnelly, N. (2009). The cost of search for multiple targets: the effects of practice and target similarity. *Journal of Experimental Psychology: Applied, 15*, 125-139.

Menneer, T, Donnelly, N, Godwin, H. J. and Cave, K. R. (2010) High or low target prevalence increases the dual-target cost in visual search. *Journal of Experimental Psychology: Applied*, 16, (2), 133-144.

Miranda A, Palmer E. (2014) Intrinsic motivation and attentional capture from gamelike features in a visual search task. Behav Res Methods, 46(1):159–72. doi: 10.3758/s13428-013-0357-7.

Mitroff, S. R., & Biggs, A. T. (2014). The Ultra-Rare-Item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science, 25*(1), 284-289. DOI: 10.1177/0956797613504221

Navalpakkam, V., Koch, C. & P. Perona, P. (2009) Homo economicus in visual search Journal of Vision, 9 (1), 16-31

Peltier, C., & Becker, M. W. (2016). Decision Processes in Visual Search as a Function of Target Prevalence. *Journal of Experimental Psychology: Human Perception and Performance. 42*, 1466-1476.

Perry N, BroedersM, deWolf C, Törnberg S, Holland R, von Karsa L (2008) European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Ann Oncol*. 19(4):614-622.

Philpotts LE (2009). Can Computer-aided Detection Be Detrimental to Mammographic Interpretation? *Radiology, 253*:17-22.

Policy on mammography screening for breast cancer version 2. City, Country: The Royal Australian and New Zealand College of Radiologists; 2014.

Posso M, Carles M, Rué M, Puig T, Bonfill X. (2016). Cost-effectiveness of double reading versus single reading of mammograms in a breast cancer screening programme. *PloS one. ;11(*7):e0159806.

Posso M, Puig T, Carles M, Rué M, Canelo-Aybar C, Bonfill X. (2017) Effectiveness and cost effectiveness of double reading in digital mammography screening: A systematic review and meta- analysis. *European Journal of Radiology. 96*:40-49.

Rich, A. N., Kunar, M. A., Van Wert, M. J., Hidalgo-Sotelo, B., Horowitz, T. S., & Wolfe, J. M. (2008). Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of Vision , 8* (15), 1-17.

Russell, N. & Kunar, M. A. (2012). Color and Spatial Cueing in Low Prevalence Visual Search. *The Quarterly Journal of Experimental Psychology, 65,* 1327-1344.

Sato M, Kawai M, Nishino Y, Shibuya D, Ohuchi N, Ishibashi T. (2014). Cost-effectiveness analysis for breast cancer screening: double reading versus single+ CAD reading. *Breast cancer. 21(5)*:532-541. 10

Schwark, J., Sandry, J., MacDonald, J., & Dolgov, I. (2012). False feedback increases detection of low-prevalence targets in visual search. Attention, Perception, and Psychophysics, 74, 1583 -1589. doi:10.3758/s13414-012-0354-4

See JE, Howe SR,Warm JS, Dember WN. (1995) Meta-analysis of the sensitivity decrement in vigilance. *Psychol Bull*. 117(2):230-249.

Taylor P & Potts HW. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. *European Journal of Cancer;44(6):*798-807.

Taylor-Phillips, Sian, Jenkinson, David J., Stinton, Chris, Wallis, Matthew G., Dunn, Janet A., Clarke, Aileen. 2018. Double reading in breast cancer screening : cohort evaluation in the CO-OPS trial. *Radiology*, 287 (3), pp. 749-757,

Taylor-Phillips, Sian, Wallis, Matthew G., Jenkinson, David J., Adekanmbi, Victor, Parsons, Helen, Dunn, Janet A., Stallard, Nigel, Szczepura, Ala, Gates, Simon, Kearins, Olive, Duncan, Alison, Hudson, Sue, Clarke, Aileen. 2016. Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection : a randomized clinical trial. *JAMA: The Journal of the American Medical Association*, 315 (18), 1956-1965.

Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2009). Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception & Psychophysics , 71* (3), 541-553.

Verster JC, Roth T. (2013) Vigilance decrement during the on-the-road driving tests: the importance of time-on-task in psychopharmacological research. *Accid Anal Prev*. 58:244-248.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2018a). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25, 35-57.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018b). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76.

Wiggins MW. (2011). Vigilance decrement during a simulated general aviation flight. *Appl Cogn Psychol*. 25(2):229-235.

Wilson R, Liston J. (2011) *Quality Assurance Guidelines for Breast Cancer Screening Radiology: NHS Breast Screening Programme Publication Number 59*. Sheffield, England: NHS Cancer Screening Programmes. NHSBS*P* publication 59.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual search. *Nature , 435*, 439-440.

Wolfe, J. M., Horowitz, T. S., Ven Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology , 136* (4), 623-638.

Wolfe, J.M., and VanWert, M.J. (2010). Varying target prevalence reveals two, dissociable decision criteria in visual search. *Current Biology, 20*, 121-124.

Table 1: Percentage of Initial and Self-Corrected Miss Errors for each Experiment. Standard Errors are reported in the Parentheses.

Condition	**Initial**		**Self-Corrected**	
	HP	**LP**	**HP**	**LP**
Experiment 1a				
Single Reading	1.9 (0.4)	15.7 (3.0)	0.4 (0.2)	7.5 (1.8)
Double Reading	1.0 (0.6)	15.8 (4.0)	0.4 (0.3)	2.5 (1.7)
Experiment 1b				
Single Reading	0.9 (0.3)	12.1 (2.2)	0.3 (0.2)	6.5 (1.6)
Double Reading	2.3 (0.8)	2.9 (1.6)	0.0 (0.0)	0.0 (0.0)
Experiment 2				
Forward - Benign	5.0 (1.1)	6.7 (1.0)	1.7 (1.1)	0.8 (0.8)
Forward - Cancer	9.2 (2.6)	15.0 (2.0)	1.7 (1.1)	7.5 (3.1)
Reverse - Benign	5.8 (1.2)	9.2 (1.0)	0.0 (0.0)	0.0 (0.0)
Reverse - Cancer	10 (2.2)	10.8 (1.1)	1.7 (1.1)	4.2 (0.8)

Table 2: Mean Correct RTs (ms) for each Experiment*. Standard Errors are reported in the parentheses.

Condition	HP	LP
Experiment 1a		
Single Reading – Present	807 (48)	1336 (102)
Single Reading – Absent	1126 (96)	959 (73)
Double Reading - Present	1055 (118)	1561 (251)
Double Reading – Absent	1256 (166)	954 (133)
Experiment 1b		
Single Reading - Present	769 (103)	1052 (70)
Single Reading – Absent	1163 (174)	967 (83)
Double Reading - Present	751 (58)	1125 (113)
Double Reading – Absent	1059 (96)	999 (98)
Experiment 2		
Present – Benign	1479 (89)	1870 (114)
Present – Cancer	1767 (102)	2210 (143)
Absent	2138 (250)	1558 (143)

* As RTs could not meaningfully be averaged across participants in the double reading conditions when participants viewed the mammograms in different rooms the data for double reading conditions in Experiments 1b and 2 are for the individual participants (not participant pairs).

Table 3: Data from individual participants (before the data were combined into paired readers) in Experiment 2 where there were multiple potential masses to search for.

Condition	**HP**	**LP**
Percentage of Miss Errors		
Benign	3.97 (1.08)	7.50 (1.41)
Cancer	8.33 (1.75)	17.08 (2.55)
Percentage of False Alarms	8.89 (2.36)	6.65 (2.00)
A Prime		
Benign	0.97 (0.01)	0.96 (0.01)
Cancer	0.95 (0.01)	0.94 (0.01)
Criteria		
Benign	-0.07 (0.08)	0.16 (0.08)
Cancer	0.06 (0.07)	0.37 (0.09)

Figure Legends

Figure 1. Multiple Decision Model with predictions for how double reading would improve LP search. Selected target items falling to the right of the criterion line would be considered a hit, else they would be considered a miss. Miss error rates could be reduced if *d'* was increased (in line with an *improved sensitivity* account, see upper right panel) or if the response bias (*c*) moved to the left and responses became more liberal (in line with a *reduced response bias* account, see lower right panel).

Figure 2. Example displays of the laboratory mammogram search where participants searched for a cancer in a mammogram. For reader clarity, the cancer in this image is in the dotted line. Please note the line did not appear in the experiment proper.

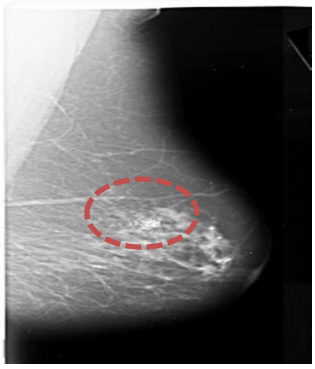
Figure 3. (a) Miss errors and (b) False alarm rates for Experiments 1a and 1b. Error bars represent the standard error.

Figure 4. (a) *A'* and (b) *c* values for Experiments 1a and 1b. Error bars represent the standard error.

Figure 5. Example displays of (a) a benign mass and (b) a cancerous mass in Experiment 2. For reader clarity, the masses are highlighted by the dotted line. Please note the line did not appear in the experiment proper.

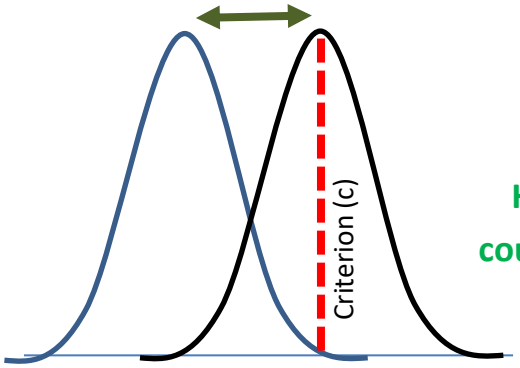
Figure 6. (a) Miss errors and (b) False alarm rates for Experiment 2. Error bars represent the standard error.

Figure 7. (a) *A'* and (b) *c* values for Experiment 2. Error bars represent the standard error.

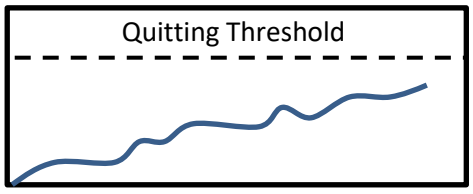


Select area for further processing

Sensitivity (d')



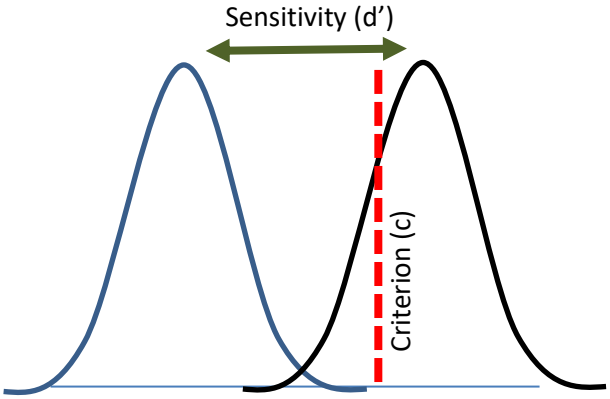
If the response criterion is exceeded then respond target present, else check quitting threshold



If the quitting threshold is exceeded then search is terminated, else repeat the process again with a new area

How double reading could reduce miss errors

INCREASE IN SENSITIVITY (D')



CHANGE IN RESPONSE BIAS (C)

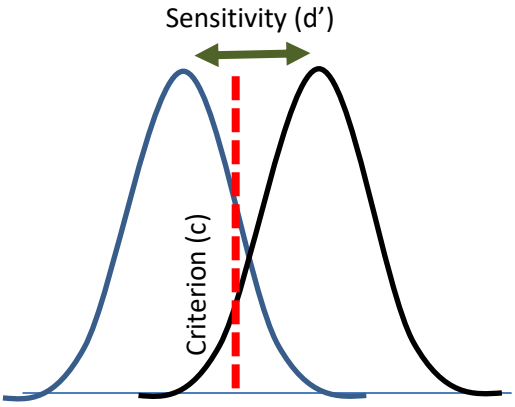


Figure 1

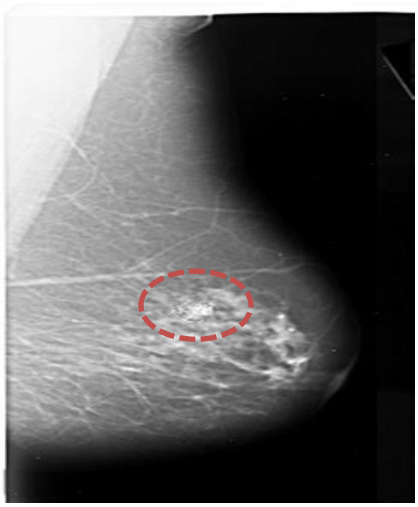
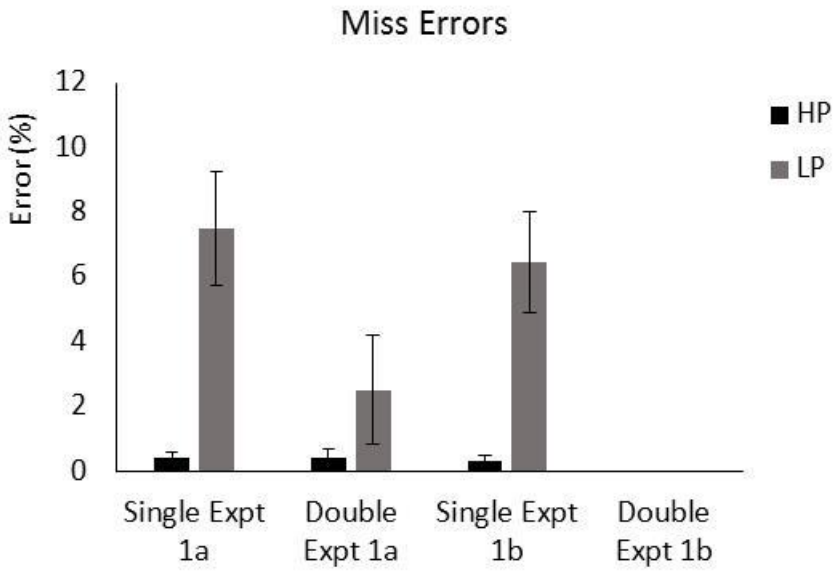


Figure 2

a)


Miss Errors
Error (%)
12
10
8
6
4
2
0
HP
LP
Single Expt 1a
Double Expt 1a
Single Expt 1b
Double Expt 1b

b)

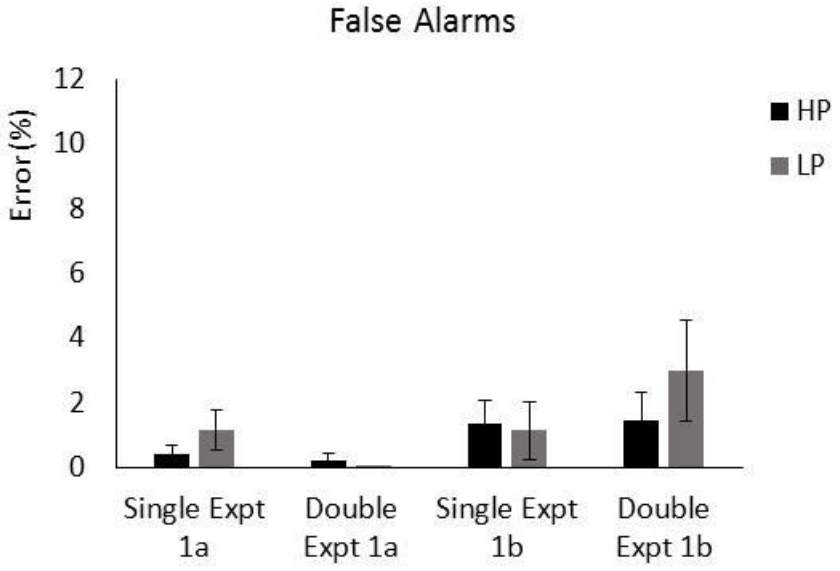
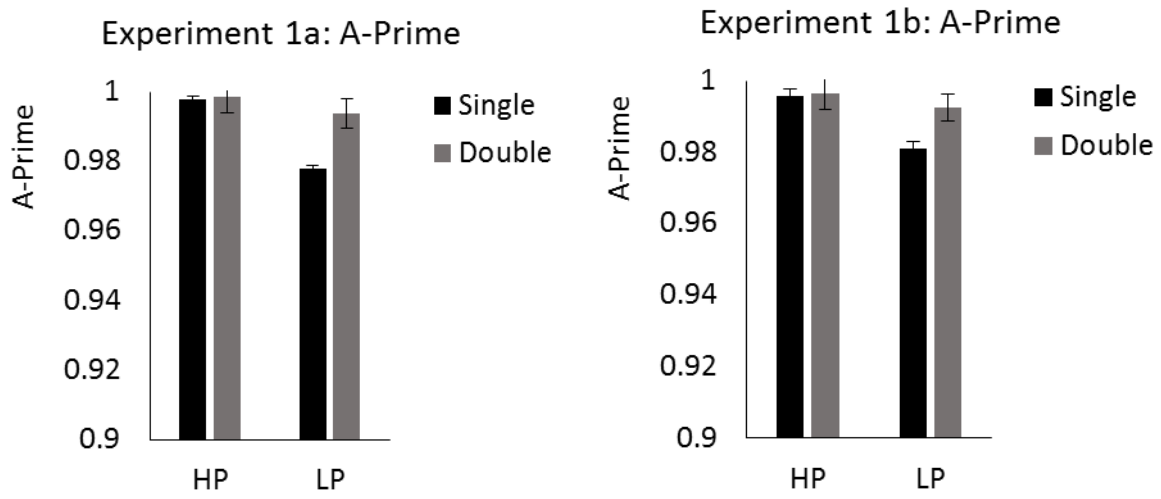

False Alarms
Error (%)
12
10
8
6
4
2
0
HP
LP
Single Expt 1a
Double Expt 1a
Single Expt 1b
Double Expt 1b

Figure 3

a)


Experiment 1a: A-Prime
A-Prime
1
0.98
0.96
0.94
0.92
0.9
HP
LP
Single
Double
Experiment 1b: A-Prime
A-Prime
1
0.98
0.96
0.94
0.92
0.9
HP
LP
Single
Double

b)

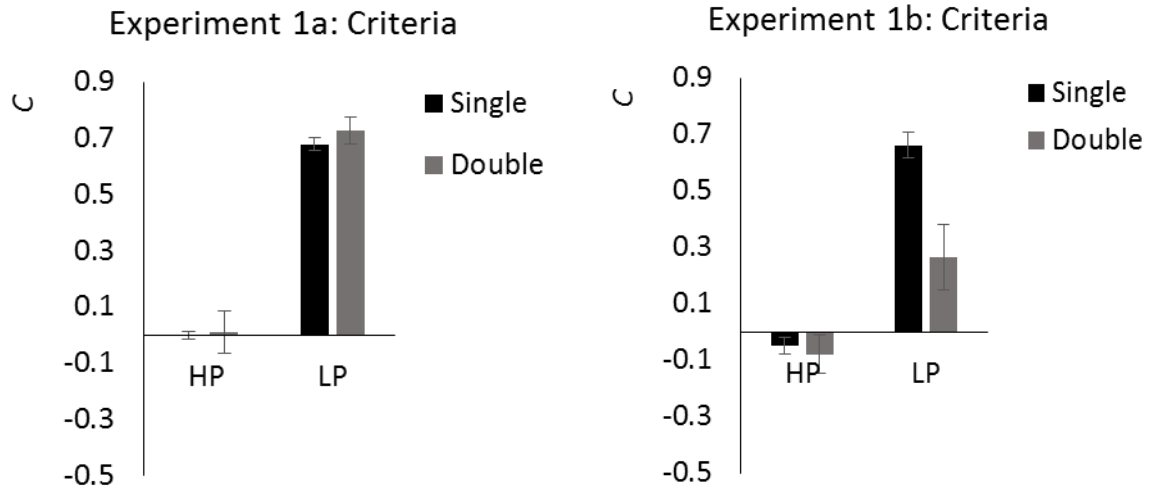
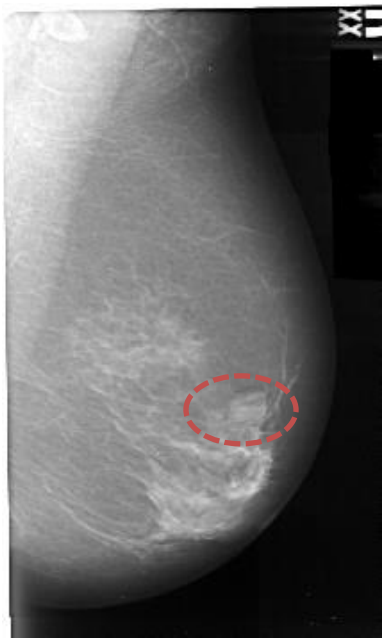

Experiment 1a: Criteria
C
0.9
0.7
0.5
0.3
0.1
-0.1
-0.3
-0.5
HP
LP
Single
Double
Experiment 1b: Criteria
C
0.9
0.7
0.5
0.3
0.1
-0.1
-0.3
-0.5
HP
LP
Single
Double

Figure 4

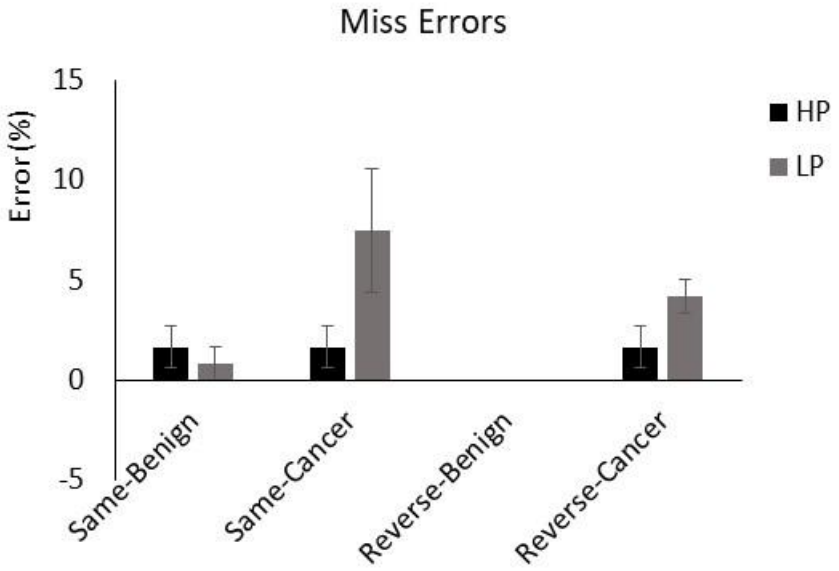
(a)



(b)

Figure 5

a)


Miss Errors
Error (%)
15
10
5
0
-5
HP
LP
Same-Benign
Same-Cancer
Reverse-Benign
Reverse-Cancer

b)

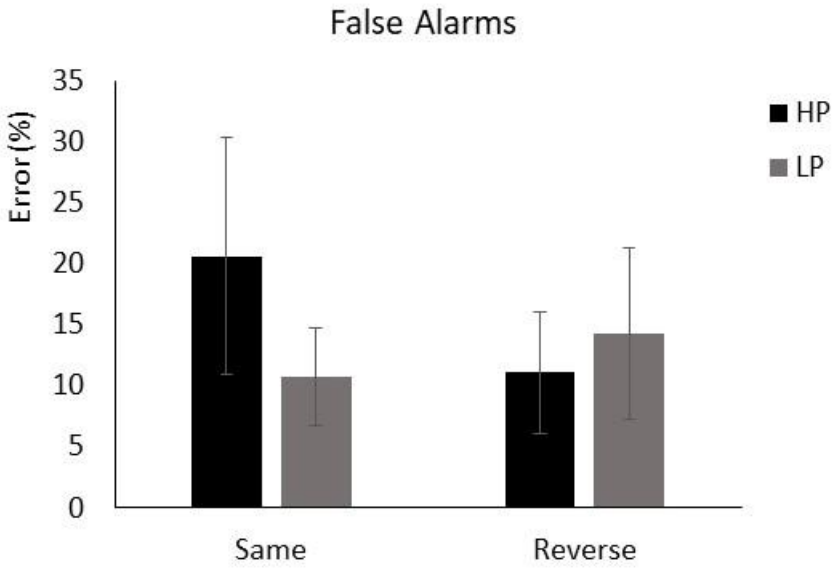
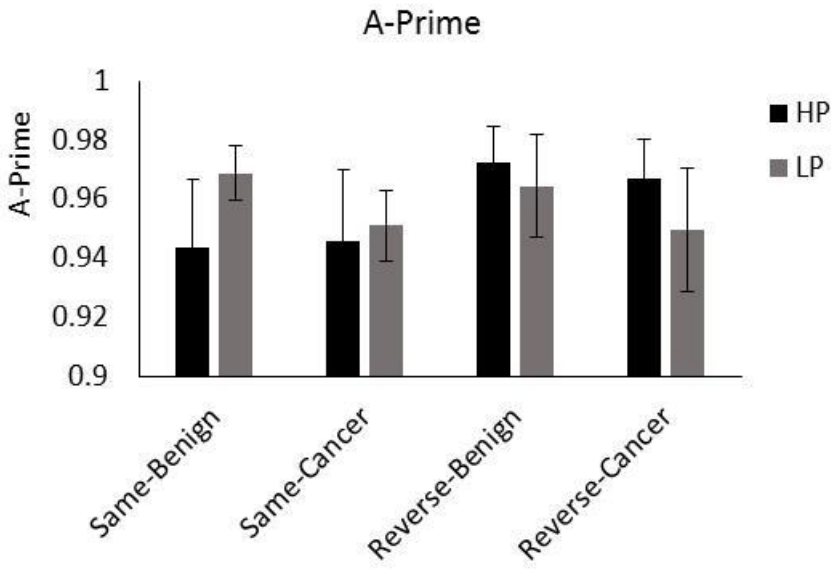

False Alarms
Error (%)
35
30
25
20
15
10
5
0
HP
LP
Same
Reverse

Figure 6

a)


A-Prime
A-Prime
1
0.98
0.96
0.94
0.92
0.9
HP
LP
Same-Benign
Same-Cancer
Reverse-Benign
Reverse-Cancer

b)

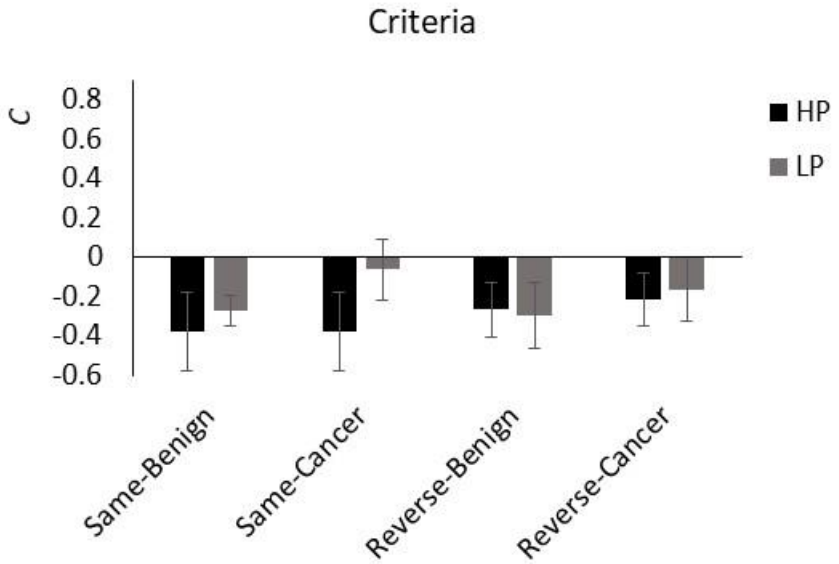

Criteria
C
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
HP
LP
Same-Benign
Same-Cancer
Reverse-Benign
Reverse-Cancer

Figure 7